A MATHEMATICAL FRAMEWORK FOR OPTIMISATION OF SEMANTIC CHUNKING IN RETRIEVAL AUGMENTED GENERATION SYSTEMS

SHORT SYNOPSIS

Submitted in partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

by

Raghvendra Kumar Jha
Registration No. **MRUCSDN012K240002**

Under the supervision of

Dr. Mamta Arora
Associate Professor

Prof. (Dr) Dipali Bansal
Dean Engineering

Department of Computer Science & Technology
School of Engineering

MANAV RACHNA vidyanatariksha
MANAV RACHNA UNIVERSITY
Declared as State Private University vide Haryana Act 26 of 2014

MANAV RACHNA UNIVERSITY

Year *2026*

MANAV RACHNA UNIVERSITY

Date. 06 May 2026

DECLARATION BY THE SCHOLAR

I hereby declare that the short synopsis entitled **A Mathematical Framework for Optimisation of Semantic Chunking in Retrieval Augmented Generation Systems** being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in the **Department of CST** under the **School of Engineering** of Manav Rachna University, Faridabad during the academic year **2024-25**.

It is the work that is expected to be carried out under the supervision of **Dr Mamta Arora** and co-supervision of **Prof (Dr) Dipali Bansal**.

Further, it is declared that the short synopsis is within the plagiarism limit and a copy of the same is attached as per the University norms.

(Raghvendra Kumar Jha)
Registration Number **MRUCSDN012K240002**

MANAV RACHNA UNIVERSITY

Date: 06 May 2026

CERTIFICATE BY SUPERVISOR(S)

This is to certify that, the short synopsis entitled **A MATHEMATICAL FRAMEWORK FOR OPTIMISATION OF SEMANTIC CHUNKING IN RETRIEVAL AUGMENTED GENERATION SYSTEMS** is an original piece of work done by research scholar **Raghvendra Kumar Jha** under the supervision of **Dr. Mamta Arora** and co-supervision of **Prof. (Dr.) Dipali Bansal** for the degree of Doctor of Philosophy in the Department of **CST** under the School of **Engineering** of Manav Rachna University, Faridabad during the academic year **2024-25.**

Further, it is certified that the short synopsis is within the plagiarism limit and a copy of the same is attached as per the University norms.

6/May/2026.

Dr. Mamta Arora
Associate Professor

Prof (Dr) Dipali Bansal
Dean Engineering

ABSTRACT

Retrieval-Augmented Generation (RAG) systems depend critically on how source documents are segmented prior to retrieval. Recent research in the field has shown that Retrieval performance in RAG systems is highly sensitive to chunk size and overlap. Existing chunking strategies, such as fixed-length, sentence-based, and embedding-driven segmentation do not explicitly model the trade-offs between semantic coherence, contextual integrity, retrieval relevance, redundancy, and computational cost, whereas semantically coherent chunks improve dense embedding quality and similarity search.

The study first defines quantitative metrics for semantic coherence and contextual integrity based on embedding similarity, information density, boundary continuity, and intra-chunk variance. It then analyzes the relationship between chunk granularity and retrieval effectiveness through both theoretical modeling and empirical evaluation using standard metrics across dense and hybrid retrievers. Building on this analysis, the research derives objective functions and constraints that formally capture trade-offs between semantic completeness, overlap control, redundancy reduction, contextual coverage, and efficiency. Standard RAG evaluation metrics will be used to validate the proposed framework. While keeping the LLM and retriever system constant, comparison of different chunking methods will be conducted based on the results obtained. This will ensure the isolation of segmentation effect. Besides this, the computational cost analysis will also be performed to determine the economic efficiency of Semantic chunking method.

This work aims to strengthen the theoretical foundations and practical reliability of semantic chunking in large-scale RAG systems.

Keywords: Retrieval-Augmented Generation (RAG), Semantic Chunking, Large Language Models, Information Retrieval, Evaluation Metrics

CONTENTS

S. No.	Description	Page No.
1	Introduction	1-2
2	Literature Review	2-6
3	Research Gap	6
4	Problem Statement	6
5	Research Objectives	6-7
6	Methodology	7-9
7	Expected outcome of the Research	9-10
8	References (Harvard Style/APA/MLA)	11-15

1. Introduction

The evolution of transformer-based architectures, particularly through the emergence of advanced models such as ChatGPT and Gemini, has marked a significant milestone in the field of Natural Language Processing (NLP). These developments have accelerated the adoption of Retrieval-Augmented Generation (RAG) particularly for enhancing the capabilities of Large Language Models (LLMs), especially within enterprises and private institutions that require secure handling of proprietary data and cannot rely on public-domain solutions. By grounding the responses of LLMs in external knowledge sources, RAG addresses inherent limitations such as knowledge cut-off and the potential for generating inaccurate or hallucinated information. As a result, extensive research has been directed not only toward the design and application of RAG systems but also toward their core components, including semantic chunking methodologies and retrieval mechanisms.

The fundamental principle of Retrieval-Augmented Generation (RAG) involves retrieving relevant information from an external knowledge base and integrating it into the input prompt provided to a large language model (LLM), thereby enabling the generation of more accurate, factual, and contextually grounded responses, as demonstrated in the retrieval-augmented generation framework proposed by Lewis et al. (2020).

Before detailing the steps involved in a Retrieval-Augmented Generation (RAG) pipeline, it is important to highlight the role of document chunking. Breaking a large document into smaller, manageable parts, known as chunking, is an important first step in the Retrieval-Augmented Generation (RAG) process. This step is necessary before moving on to indexing and retrieving information. Studies show that how we divide the text affects how accurately information can be retrieved and how good the generated responses are (Lewis et al., 2020; LlamaIndex, 2023). Recent research in the field also has shown that if chunking is done poorly, retrieval can be less precise, and the quality of the output can drop (Gao et al., 2023). Traditional methods, like dividing text into fixed sizes or using simple rules, often break the text in ways that lose the flow of meaning, making it harder for the system to understand the context.

In Semantic chunking text is split into parts by maintaining context to keep complete ideas. This approach helps improve both retrieval and the responses generated by language models.

This research focuses on understanding semantic chunking in RAG, including what it is, how it works, its benefits and challenges, how it compares to other methods, and the latest tools and developments in this area.

2. Literature Review:

This section presents a comprehensive review of the research landscape relevant to semantic chunking and its optimization within Retrieval-Augmented Generation (RAG) systems. The discussion traces the evolution from classical text segmentation and probabilistic information retrieval to dense semantic representations, transformer-based language models, and modern RAG architectures. Across these developments, a consistent limitation emerges; although chunking critically influences retrieval accuracy, generation faithfulness, and computational efficiency, it remains largely heuristic and lacks a principled mathematical optimization framework.

2.1 Classical Text Segmentation and Probabilistic Retrieval

Early work on text segmentation treated chunking as a discourse cohesion problem. Hearst introduced **TextTiling**, an unsupervised algorithm that segments documents into multi-paragraph topical units using lexical similarity patterns [1] **.** This work demonstrated that segmentation quality directly affects downstream information access but relied exclusively on surface-level term statistics.

In parallel, foundational information retrieval models formalized relevance estimation. Salton and Buckley proposed TF–IDF weighting to quantify term importance across documents [2]**,** while Robertson and Zaragoza extended probabilistic relevance modeling through the BM25 framework [3]**.** Ponte and Croft reframed retrieval as a language modeling problem by estimating the likelihood of a query being generated from a document model [4]**.** Although theoretically grounded, these approaches implicitly assumed that documents or passages were already coherent semantic units.

Subsequent work on event-centric information retrieval attempted to incorporate higher-level semantic structure beyond lexical overlap [5]. However, classical IR and segmentation methods remained constrained by term-based similarity and were unable to robustly capture semantic continuity across chunk boundaries.

2.2 Distributed Semantic Representations and Early Semantic Chunking

The emergence of distributed semantic representations marked a shift from symbolic to continuous modeling of meaning. Mikolov et al. demonstrated that neural word embeddings encode syntactic and semantic regularities in vector space [6]. Pennington et al. further advanced this paradigm through GloVe, which combines global co-occurrence statistics with local context learning [7]. These representations enabled semantic similarity computation beyond exact word matching.

Alemi and Ginely were among the first to explicitly apply semantic embeddings to text segmentation, proposing similarity-based boundary detection using distributed vectors [8]. The introduction of Sentence-BERT enabled sentence-level embeddings optimized for similarity search[9], forming the basis of modern dense retrieval pipelines. Despite these advances, segmentation strategies largely remained heuristic, relying on fixed window sizes or rule-based splitting.

2.3 Transformer Models and Context Window Constraints

The Transformer architecture enabled global self-attention and deep contextual representation learning [10]. This architecture underpinned large-scale pretraining models such as BERT [11]and T5 [12], which demonstrated strong transfer learning across a wide range of NLP tasks.

As model scale increased, large language models such as GPT-3 [13], LLaMA-2 [14], and GPT-4 [15] exhibited strong reasoning and generation capabilities. However, all transformer-based models are constrained by finite context windows. Liu et al. empirically demonstrated the *Lost-in-the-Middle* phenomenon, showing that language models struggle to reliably access relevant information embedded within long or fragmented contexts [16].

Efforts to mitigate this limitation include retrieval-enhanced architectures such as RETRO [17] and knowledge-enhanced pretraining approaches such as Ernie 3.0 [18]**.** Nevertheless, these systems still depend on externally defined chunk boundaries, highlighting the unresolved role of segmentation in long-context reasoning.

2.4 Retrieval-Augmented Generation Architectures

Retrieval-Augmented Generation was formalized by Lewis et al., who combined parametric language models with non-parametric retrieval over external knowledge sources [19]**.**

REALM introduced joint optimization of retrieval and language modelling during pretraining [20], while ColBERT proposed late interaction mechanisms enabling fine-grained token-level similarity matching [21]**.**

Further extensions addressed reasoning and robustness. Generation-augmented retrieval improved evidence selection [22]**,** while multi-hop retrieval and reasoning were explored for complex question answering [23]**, [24]**. Robustness to irrelevant context was studied [25]**,** and Active RAG introduced dynamic retrieval control mechanisms [26].

Advanced frameworks such as Re2G [27]**,** REPLUG [28]**,** In-Context RAG [29]**,** and Self-RAG [30] further refined retrieval–generation coupling. Survey studies emphasize that chunking strategy is a critical yet under-theorized hyperparameter across nearly all RAG systems [31]**.**

2.5 Dense Retrieval and Vector Search Optimization

Efficient semantic retrieval relies on scalable nearest-neighbor search. FAISS enabled billion-scale vector similarity search using GPUs [32], while HNSW graphs provided efficient approximate nearest-neighbor retrieval with logarithmic complexity [33]. These techniques underpin modern vector databases used in RAG pipelines.

Dense retriever training methods such as DPR [34]**,** ANCE [35]**,** topic-aware negative sampling [36], and latent retrieval [37] improved embedding discrimination. kNN-LM

demonstrated that external vector memory can significantly enhance language model generalization [38].

However, dense retrieval performance implicitly depends on the semantic coherence of chunks. Misaligned or fragmented chunks degrade embedding quality and retrieval precision, yet this dependency is rarely formalized in retrieval objectives.

2.6 Learned and Hierarchical Segmentation Models

Several works explored segmentation as a learnable task. Text segmentation was formulated as supervised learning [39], while attention-based neural segmentation models were introduced [40]. Cross-segment attention enabled modelling of global coherence across segment boundaries [41].

Earlier machine comprehension systems emphasized the importance of selecting semantically coherent spans [42], [43][44] indirectly motivating improved chunking strategies.

Recent RAG-focused studies explicitly target segmentation. Chunk filtering strategies were proposed to reduce retrieval noise [45], while semantic augmentation techniques improved answer grounding [46]. Hierarchical chunking strategies were explored to balance granularity and efficiency [47]. Empirical analyses examined trade-offs between chunk size, computational cost, and retrieval accuracy [48], [49]

2.7 Evaluation Metrics, Hallucination, and Trustworthiness

Evaluation of RAG systems traditionally relies on lexical overlap metrics such as BLEU [50] and ROUGE [51]. Semantic-aware evaluation was introduced through contextual similarity metrics [52], while retrieval-specific evaluation frameworks assessed grounding and faithfulness [53].

Hallucination detection has gained increasing attention. Empirical studies analyzed hallucination patterns in RAG pipelines , while retrieval noise was shown to significantly amplify hallucinations [54]. Search-augmented instruction learning further emphasized that coherent and semantically optimized chunks are essential for stable instruction following [55].

2.8 Practical Frameworks and Tooling

Retrieval-augmented question answering with generative readers was demonstrated through end-to-end architectures [56]. Trust and reliability of RAG prompts were analyzed using benchmark datasets . Widely used open-source frameworks provide practical implementations of recursive and semantic chunking strategies [57] [58]. However, these systems mainly depend on heuristic settings and do not use structured optimization techniques.

3. Research Gap

Earlier researches have studied the different aspects of chunking (semantic coherence, retrieval likelihood, computational cost and generation accuracy) in isolation. There have been a lack of unified framework which encapsulates all the aspect within. Without a comprehensive framework that considers these trade-offs it is hard to consistently reproduce results; scale RAG systems efficiently or systematically improve the chunking quality. This research aims to address this gap.

4. Problem Statement

To develop a mathematical framework for the optimization of semantic chunking in Retrieval-Augmented Generation systems that maximizes retrieval relevance and generation quality while minimizing semantic fragmentation and computational overhead.

5. Research Objectives

Based on the research gap identified as above (para 4) the objectives of this research are as follows:

i. To define quantitative metrics for semantic coherence and contextual integrity.

ii. To analyse chunk granularity and its effectiveness on retrieval performance.

iii. To optimize semantic chunking algorithms design.

iv. To validate proposed experimental framework with existing chunking strategies.

v. To analyze the impact of optimized chunking on generation quality.

6. Methodology

The methodology adopted in this research is designed to systematically develop, analyze, and validate a unified mathematical framework for optimizing semantic chunking in Retrieval-Augmented Generation (RAG) systems. The methodology integrates theoretical modeling, algorithmic design, and empirical evaluation to ensure both rigor and practical relevance.

6.1 Evaluation of Existing Semantic Chunking Techniques

A comprehensive analysis of state-of-the-art semantic chunking techniques will be conducted, including fixed-length, sentence-based, embedding-based, and adaptive chunking methods. This review will examine their underlying design assumptions, strengths, limitations, and suitability for RAG pipelines. Attention will be given to how existing approaches handle semantic coherence, chunk boundaries, overlap strategies, and computational constraints. The analysis will also identify gaps in current evaluation practices, motivating the need for a unified mathematical formulation.

6.2 Dataset Selection, Curation, and Preparation

Datasets will be carefully selected and prepared to support both the theoretical study and practical evaluation of semantic chunking strategies. Datasets for this research will be drawn from public repositories like Hugging Face and Kaggle, with selection guided by factors such as document length, semantic complexity, and coverage across different domains to ensure a thorough evaluation. A detailed review of previous RAG and semantic chunking studies will help identify the datasets, methods, and experimental setups commonly used, ensuring that results are comparable and reproducible. If existing datasets fall short, a custom dataset will

be created to address more challenging aspects, including ambiguous semantic boundaries, variable information density, and long-range contextual dependencies, which are essential for rigorously testing RAG systems.

6.3 Mathematical Model Formalization of Semantic Chunking

A unified mathematical framework will be developed to model semantic chunking as a constrained optimization problem. Formal definitions will be introduced for semantic coherence, contextual integrity, retrieval relevance, and computational cost using embedding-based representations and information-theoretic measures. Objective functions and constraints will be derived to capture trade-offs between chunk granularity, overlap, semantic completeness, and efficiency, providing a principled basis for chunk boundary determination.

6.4 Optimization-Based Semantic Chunking Algorithm Design

Based on the formulated mathematical framework, an optimization-driven semantic chunking algorithm will be proposed. The algorithm will employ suitable optimization techniques such as constrained optimization, dynamic programming, or graph-based segmentation to generate semantically coherent chunks while satisfying retrieval and computational constraints. The proposed approach will be compared conceptually against heuristic and embedding-based chunking strategies to highlight its theoretical advantages.

6.5 Experimental Evaluation and Validation

The proposed chunking framework will be empirically evaluated within standardized RAG pipelines under controlled experimental settings. Retrieval performance will be assessed using metrics such as Recall@k, Mean Reciprocal Rank (MRR), and nDCG, while generation quality will be evaluated using ROUGE (Precision, Recall, and F1) and RAGAS-based faithfulness and relevance metrics. Experiments will be designed to isolate the impact of chunking decisions by keeping retrievers and language models constant across comparisons.

6.6 Computational Cost and Efficiency Analysis

A detailed computational cost analysis will be conducted to evaluate the efficiency of the proposed optimization-based chunking method. Metrics such as chunk generation time, indexing overhead, memory consumption, and retrieval latency will be measured and compared against baseline chunking techniques. This analysis will quantify the trade-offs between improved semantic and retrieval performance and the associated computational overhead.

6.7 Tools, Libraries, and Experimental Frameworks

The research will utilize established tools and frameworks to ensure reproducibility and experimental rigor:

- **NLP and Embedding Libraries:** SpaCy, NLTK, Sentence Transformers, and related text-processing frameworks.
- **Semantic Chunking and RAG Frameworks:** LangChain semantic chunking utilities, LlamaIndex-based chunking tools, and custom implementations derived from the proposed framework.
- **Evaluation Frameworks:** ROUGE, RAGAS, and additional retrieval and generation metrics as required.
- **Implementation Environment:** Python-based experimental pipelines integrating dense and hybrid retrievers with state-of-the-art language models.

7. Expected Outcome

The proposed research is expected to result in formulation of mathematical framework for the optimization of semantic chunking in Retrieval-Augmented Generation (RAG) systems. Quantitative metric for determining semantic coherence and contextual integrity will be defined. Effect of various factors of chunking on retrieval performance and cost effectiveness of semantic chunking algorithm will also be analysed. The study aims to strengthen the theory,

methodology and practical impact of semantic chunking in production Retrieval Augmented Generation systems.

8. References

[1] M. A. Hearst, "TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages."

[2] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, pp. 513–523, 1988, doi: 10.1016/0306-4573(88)90021-0.

[3] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, pp. 333–389, 2009, doi: 10.1561/1500000019.

[4] J. M. Ponte and W. B. Croft, "A Language Modeling Approach to Information Retrieval," in *SIGIR 1998 - Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, Aug. 1998, pp. 275–281. doi: 10.1145/290941.291008.

[5] G. Glavaš and J. Šnajder, "Event-Centered Information Retrieval Using Kernels on Event Graphs," https://aclanthology.org/W13-5001/.

[6] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality."

[7] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2014, pp. 1532–1543. doi: 10.3115/v1/d14-1162.

[8] A. A. Alemi and P. Ginsparg, "Text Segmentation based on Semantic Word Embeddings," Mar. 2015.

[9] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, Association for Computational Linguistics, 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

[10] A. Vaswani *et al.*, "Attention Is All You Need."

[11] J. Devlin, M.-W. Chang, K. Lee, K. T. Google, and A. I. Language, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." [Online]. Available: https://github.com/tensorflow/tensor2tensor

[12] C. Raffel *et al.*, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," https://jmlr.org/papers/v21/20-074.html.

[13] T. B. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, 2020.

[14] H. Touvron *et al.*, "Llama 2: Open Foundation and Fine-Tuned Chat Models," Jul. 2023.

[15] OpenAI *et al.*, "GPT-4 Technical Report," Mar. 2024.

[16] N. F. Liu *et al.*, "Lost in the Middle: How Language Models Use Long Contexts," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 157–173, 2024, doi: 10.1162/tacl_a_00638.

[17] S. Borgeaud *et al.*, "Improving Language Models by Retrieving from Trillions of Tokens," in *Proceedings of Machine Learning Research*, ML Research Press, 2022, pp. 2206–2240.

[18] Y. Sun *et al.*, “ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation,” Jul. 2021.
[19] P. Lewis *et al.*, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.
[20] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “Retrieval Augmented Language Model Pre-Training,” https://proceedings.mlr.press/v119/guu20a.html.
[21] O. Khattab and M. Zaharia, “ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT,” in *SIGIR 2020 - Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, Jul. 2020, pp. 39–48. doi: 10.1145/3397271.3401075.
[22] Y. Mao *et al.*, “Generation-augmented retrieval for open-domain question answering,” in *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2021, pp. 4089–4100. doi: 10.18653/v1/2021.acl-long.316.
[23] H. Trivedi, N. Balasubramanian, T. Khot, and A. Sabharwal, “Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions,” in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics (ACL), 2023, pp. 10014–10037. doi: 10.18653/v1/2023.acl-long.557.
[24] H. Liu *et al.*, “HopRAG: Multi-Hop Reasoning for Logic-Aware Retrieval-Augmented Generation,” Association for Computational Linguistics (ACL), Aug. 2025, pp. 1897–1913. doi: 10.18653/v1/2025.findings-acl.97.
[25] O. Yoran, T. Wolfson, O. Ram, and J. Berant, “MAKING RETRIEVAL-AUGMENTED LANGUAGE MODELS ROBUST TO IRRELEVANT CONTEXT,” in *12th International Conference on Learning Representations, ICLR 2024*, International Conference on Learning Representations, ICLR, 2024.
[26] Z. Jiang *et al.*, “Active Retrieval Augmented Generation,” in *EMNLP 2023 - 2023 Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2023. doi: 10.18653/v1/2023.emnlp-main.495.
[27] R. Ofori-Boateng, M. Aceves-Martins, N. Wirantuga, and C. F. Moreno-García, “A Zero-Shot Monolingual Dual Stage Information Retrieval System for Spanish Biomedical Systematic Literature Reviews,” Long Papers. [Online]. Available: https://spacy.io/models/es
[28] W. Shi *et al.*, “REPLUG: Retrieval-Augmented Black-Box Language Models,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, Association for Computational Linguistics (ACL), 2024, pp. 8364–8377. doi: 10.18653/v1/2024.naacl-long.463.
[29] O. Ram *et al.*, “In-Context Retrieval-Augmented Language Models,” *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 1316–1331, 2023, doi: 10.1162/tacl_a_00605.
[30] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, “SELF-RAG: LEARNING TO RETRIEVE, GENERATE, AND CRITIQUE THROUGH SELF-REFLECTION,” in

12th International Conference on Learning Representations, ICLR 2024, International Conference on Learning Representations, ICLR, 2024.
[31] Y. Gao *et al.*, "Retrieval-Augmented Generation for Large Language Models: A Survey," Mar. 2024, doi: https://doi.org/10.48550/arXiv.2312.10997.
[32] M. Douze *et al.*, "The Faiss library," Oct. 2025, [Online]. Available: http://arxiv.org/abs/2401.08281
[33] Y. A. Malkov and D. A. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, pp. 824–836, Apr. 2020, doi: 10.1109/TPAMI.2018.2889473.
[34] V. Karpukhin *et al.*, "Dense passage retrieval for open-domain question answering," in *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2020, pp. 6769–6781. doi: 10.18653/v1/2020.emnlp-main.550.
[35] "Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval," https://iclr.cc/virtual/2021/poster/2673.
[36] S. Hofstätter, S. C. Lin, J. H. Yang, J. Lin, and A. Hanbury, "Efficiently Teaching an Effective Dense Retriever with Balanced Topic Aware Sampling," in *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, Jul. 2021, pp. 113–122. doi: 10.1145/3404835.3462891.
[37] K. Lee, M. W. Chang, and K. Toutanova, "Latent retrieval for weakly supervised open domain question answering," in *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2020, pp. 6086–6096. doi: 10.18653/v1/p19-1612.
[38] U. Khandelwal, O. Levy, D. Jurafsky, L. Zettlemoyer, and M. Lewis, "GENERALIZATION THROUGH MEMORIZATION: NEAREST NEIGHBOR LANGUAGE MODELS," in *8th International Conference on Learning Representations, ICLR 2020*, International Conference on Learning Representations, ICLR, 2020.
[39] O. Koshorek, A. Cohen, N. Mor, M. Rotman, and J. Berant, "Text segmentation as a supervised learning task," in *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2018, pp. 469–473. doi: 10.18653/v1/n18-2075.
[40] P. Badjatiya, L. J. Kurisinkel, M. Gupta, and V. Varma, "Attention-based neural text segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer Verlag, 2018, pp. 180–193. doi: 10.1007/978-3-319-76941-7_14.
[41] M. Lukasik, B. Dadachev, K. Papineni, and G. Simões, "Text Segmentation by Cross Segment Attention."
[42] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading Wikipedia to answer open-domain questions," in *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, Association for Computational Linguistics (ACL), 2017, pp. 1870–1879. doi: 10.18653/v1/P17-1171.

[43] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, “BI-DIRECTIONAL ATTENTION FLOW FOR MACHINE COMPREHENSION,” in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, International Conference on Learning Representations, ICLR, 2017.
[44] D. Golub, P.-S. Huang, X. He, and L. Deng, “Two-Stage Synthesis Networks for Transfer Learning in Machine Comprehension.” [Online]. Available: https://github.com/
[45] I. S. Singh *et al.*, “ChunkRAG: Novel LLM-Chunk Filtering Method for RAG Systems,” Apr. 2025.
[46] K. Zhong, B. Suleiman, A. Erradi, and S. Chen, “SemRAG: Semantic Knowledge-Augmented RAG for Improved Question-Answering,” Jul. 2025.
[47] H. T. Nguyen, T. D. Nguyen, and V. H. Nguyen, “Enhancing Retrieval Augmented Generation with Hierarchical Text Segmentation Chunking,” in *Communications in Computer and Information Science*, Springer Science and Business Media Deutschland GmbH, 2025, pp. 209–220. doi: 10.1007/978-981-96-4288-5_17.
[48] J. Hladěna, K. Šteflovič, P. Čech, K. Štekerová, and A. Žváčková, “The Effect of Chunk Size on the RAG Performance,” in *Lecture Notes in Networks and Systems*, Springer Science and Business Media Deutschland GmbH, 2025, pp. 317–326. doi: 10.1007/978-3-032-00712-4_21.
[49] R. Qu, R. Tu, and F. Bao, “Is Semantic Chunking Worth the Computational Cost?,” Oct. 2024.
[50] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “BLEU,” in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL ’02*, Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.
[51] C.-Y. Lin, “ROUGE: A Package for Automatic Evaluation of Summaries,” https://aclanthology.org/W04-1013/.
[52] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTSCORE: EVALUATING TEXT GENERATION WITH BERT,” in *8th International Conference on Learning Representations, ICLR 2020*, International Conference on Learning Representations, ICLR, 2020.
[53] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” in *EACL 2024 - 18th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of System Demonstrations*, Association for Computational Linguistics (ACL), 2024, pp. 150–158. doi: 10.18653/v1/2024.eacl-demo.16.
[54] F. Cuconasu *et al.*, “The Power of Noise: Redefining Retrieval for RAG Systems,” in *SIGIR 2024 - Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, Jul. 2024, pp. 719–729. doi: 10.1145/3626772.3657834.
[55] H. Luo *et al.*, “SAIL: Search-Augmented Instruction Learning,” Jun. 2023.
[56] G. Izacard and E. Grave, “Leveraging passage retrieval with generative models for open domain question answering,” in *EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference*, Association for Computational Linguistics (ACL), 2021, pp. 874–880. doi: 10.18653/v1/2021.eacl-main.74.
[57] “LangChain Reference Docs,” https://reference.langchain.com/.

[58] “Welcome to LlamaIndex 🦙 !ℑ,” https://llama-index.readthedocs.io/zh/stable/.

Refal

MANAV RACHNA UNIVERSITY

No. MRU/SVLRC/LIB: .....20....
06 May 2026

Swami Vivekananda Library and Resource Centre
Manav Rachna University

SIMILARITY VERIFICATION CERTIFICATE

Title: A MATHEMATICAL FRAMEWORK FOR OPTIMISATION OF SEMANTIC CHUNKING IN RETRIEVAL AUGMENTED GENERATION SYSTEMS

Researcher: RAGHVENDRA KUMAR JHA Registration Number: MRUCSDW012K240002

Supervisor: Dr MAMTA ARORA Department: CST

This is to certify that the above mentioned Short Synopsis/Thesis was scanned for similarity index through Drill Bit Software on 16-04-2026 with submission ID 5497986 and the similarity index was found to be 02 %. The complete report running into 16 Pages, 4134 Word count with File Size 1071 KB is being forwarded to the Research Supervisor/DRC/DRB for review.

The complete report of the above Short Synopsis/~~Thesis~~ has been reviewed and the similarity index including AI-based writing tools are below the accepted norms as per the University Grants Commission (Promotion of Academic Integrity and Prevention of Plagiarism in Higher Educational Institutions) Regulations 2018. The Short Synopsis/~~Thesis~~ may be considered for further processing.

06 May 26
Research Scholar

06/May/2026
Research Supervisor

Swami Vivekananda Library
6/5/2026
Library

Excluding matches less than 14 words

Excluding matches less than 14 words

6/May/2026.

6/5/2026

DrillBit

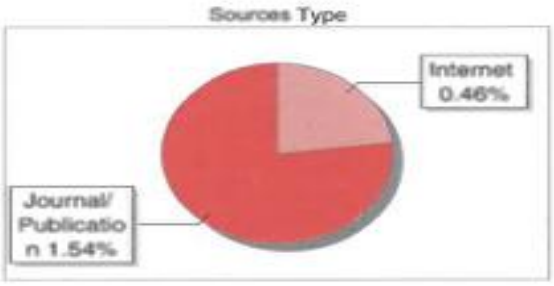
The Report is Generated by DrillBit Plagiarism Detection Software

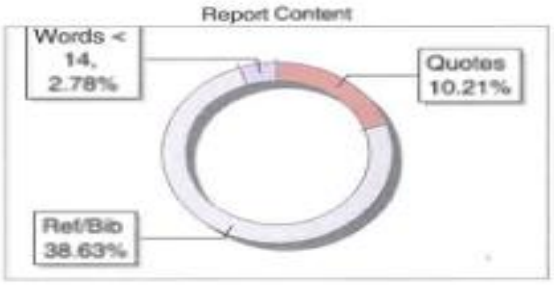
Submission Information

Author Name	Raghvendra Kumar Jha
Title	Short Synopsis
Paper/Submission ID	5497986
Submitted by	netrapal.library@mru.edu.in
Submission Date	2026-04-16 11:35:50
Total Pages, Total Words	16, 4134
Document type	Synopsis

Result Information

Similarity 2 %





Exclude Information

Quotes	Excluded
References/Bibliography	Excluded
Source: Excluded < 14 Words	Excluded
Excluded Source	0 %
Excluded Phrases	Not Excluded

Database Selection

Language	English
Student Papers	Yes
Journals & publishers	Yes
Internet or Web	Yes
Institution Repository	Yes

A Unique QR Code use to View/Download/Share Pdf File

Excluding matches
less than 14 words

6/May/2026

DrillBit

DrillBit Similarity Report

2	2	A
SIMILARITY %	MATCHED SOURCES	GRADE

A-Satisfactory (0-10%)
B-Upgrade (11-40%)
C-Poor (41-60%)
D-Unacceptable (61-100%)

LOCATION	MATCHED DOMAIN	%	SOURCE TYPE
5	arxiv.org	2	Publication
18	pmc.ncbi.nlm.nih.gov	<1	Internet Data

6/May/2026

A MATHEMATICAL FRAMEWORK FOR OPTIMISATION OF SEMANTIC CHUNKING IN RETRIEVAL AUGMENTED GENERATION SYSTEMS

SHORT SYNOPSIS

by

Raghvendra Kumar Jha
Registration No. **MRUCSDN012K240002**

Under the supervision of

Dr. Mamta Arora
Associate Professor

Prof. (Dr) Dipali Bansal
Dean Engineering

MANAV RACHNA UNIVERSITY

Year *2026*

ABSTRACT

DrillBit

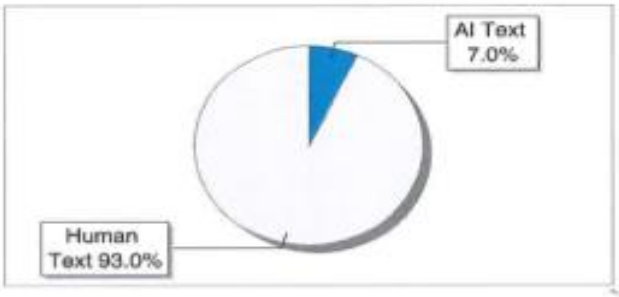
The Report is Generated by DrillBit AI Content Detection Software

Submission Information

Author Name	Raghvendra Kumar Jha
Title	Short Synopsis
Paper--Submission ID	5497986
Submitted By	netrapal.library@mru.edu.in
Submission Date	2026-04-16 11:35:50
Total [Pages,Sentences,Words]	[16, 130, 2475]
Document type	Synopsis
Language	English

Result Information

AI Text: **7 %**



Disclaimer:

* The content detection system employed here is powered by artificial intelligence (AI) technology.
* Its not always accurate and only help to author identify text that might be prepared by a AI tool.
* It is designed to assist in identifying & moderating content that may violate community guidelines/legal regulations, it may not be perfect.